Geospatial Data Disclosure Avoidance and the Census

Select Topics in International Censuses[1]

By Sam Dupre, Lindsay Spell, and Paul Jung

Released March 2022

INTRODUCTION

Geocoded data are becoming available at increasingly finer resolutions, while attacker capabilities in linking geocoded data to specific individuals or groups continue to grow. The number of identifying variables necessary for a data breach to occur is reduced dramatically when geographic location is known at a precise level, especially where population density is low. The development of geocoded data sources presents new concerns for data confidentiality breaches, as traditional disclosure avoidance mechanisms may ignore spatial characteristics (Buron and Fontaine, 2018).

While geographic characteristics are technically just another variable, there are certain aspects that make them worthy of further consideration.

National statistical offices (NSOs) are currently in the challenging position of having to disseminate census data to users, while also preserving spatial relationships in the data and taking spatial data features into account in statistical disclosure control. It is important to note that the sensitivity of variables and populations varies widely between national contexts; there is no one-size-fits-all approach to disclosure avoidance.

This technical note discusses key concepts in geospatial disclosure avoidance and includes a process overview with an example case for implementation. This note builds on the material presented in "Disclosure Avoidance and the Census" (U.S. Census Bureau, 2020). Refer to Box 1 for a summary of that guidance.

[1] This technical note is part of a series on Select Topics in International Censuses (STIC) that explore matters of interest to the international statistical community. The U.S. Census Bureau helps countries improve their national statistical systems by engaging in capacity building to enhance statistical competencies in sustainable ways. Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau.

Box 1.

General Disclosure Avoidance

Statutory Requirement

While statutory requirements to protect respondent data privacy vary from country to country, these protections are critical in supporting public trust and willingness to participate in censuses and survey efforts.

Avoidance Steps

1. Eliminate Personally Identifiable Information (PII).
2. Identify sensitive records, cells, and categories.
3. Address the risks.
4. Check results.
5. Conduct internal attack studies.

Types

- Identity disclosure—respondent identity is directly linked to a disseminated data record.
- Attribute disclosure—values in disseminated data disclose other attributes of an individual.
- Inferential disclosure—disseminated data are used to infer values for specific respondents based on statistical properties of the released data.

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

United States Agency for International Development
www.usaid.gov

United Nations Population Fund
www.unfpa.org

KEY CONCEPTS

Small-Area Estimation Concerns

Many NSOs release small-area population datasets to provide a fine spatial characteristic of the population distribution, even down to the collection unit/census block. However, the release of fine-resolution data may increase risk of identity disclosure. Since small-area data are likely to have lower cell values, there is a higher chance of reverse geocoding occurring (refer to Box 2) with subsequent identity disclosure by using the precise spatial attributes and the characteristics of the local population distribution within the area.

Nested Hierarchies

Regarding nested data, it is straightforward to assess differencing attack risk (refer to Box 2), but nonnested data create more complicated situations.

If the geospatial datasets are constructed with nested hierarchies, identity disclosure can be prevented by data suppression on census areas with a low valued cell. There is no overlapping area between the same-level census areas, so a differencing attack is not likely to be successful.

However, nonnested data create a high risk of differencing attacks and identity disclosure. For example, when a nonnested zone is placed within a census area, differencing a nonnested zone and the surrounding census area creates a limited number of households in the differenced area that can be potentially identified as the respondents' locations. If the differenced area has a low value attribute and sparse population distribution, there is a high chance that respondents' locations and identities could be disclosed. A similar differencing attack can also occur when a nonnested zone or gridded cell is placed across two census areas.

Microdata and Aggregate Data

NSOs release two types of data to the public: microdata and aggregate data (U.S. Census Bureau, 2020). Each form may be released with spatial information in different ways. Additional details of the procedures associated with each form are covered later in this document.

Microdata are for individual respondents that have been anonymized and may be released with a spatial attribute indicating the location of the individual respondents. When the microdata are released with coordinates, geomasking (discussed later in this guide) anonymizes their precise locations and prevents identity disclosure by infusing synthetic errors on the coordinates.

Aggregate data occur when individual cases are spatially aggregated using administrative or statistical units (e.g., districts or gridded cells) to provide area-based location information (Figure 1). Since the aggregated data with spatial attributes provide coarsened location information

Box 1.—Con.

Principles

- Release samples from larger/high-density populations.
- Reduce variability below a threshold.
- Suppress data where unique cases are visible.
- Maintain original data relationships/structure where possible.

Source: U.S. Census Bureau, 2020.

Box 2.

Some Forms of Attacks

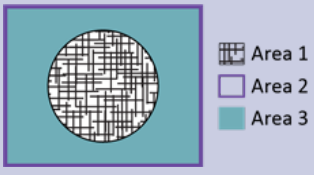
Reverse Geocoding

Geocoding is the process to convert a text-based address or place name into geographical coordinates to identify an exact location of the survey data.

Reverse geocoding occurs when the street address for a location published in paper or digital format is determined. Street addresses can then be a powerful key for linking the data associated with those coordinates to specific and identifiable information from other sources. The locations of the respondents can be inferred and reverse geocoded if a census area has low-value cells leading to possible disclosure.

Differencing Attacks

These attacks occur when data released at different levels of geography can be combined to reconstruct data at a smaller level or identify the location of an observation. For example, in the figure below, Area 1 is a subset of Area 2. A table could then be created that represents the statistics for Area 3 that could fall below confidentiality standards.



If geographical differentiation lets a geography be created with fewer than the minimum threshold of respondents, then a data breach has occurred.

Source: U.S. Census Bureau, 2021.

of the respondents, their privacy can be protected by data suppression on the areas with low values.

Geomasking

Geomasking is a geospatial process that improves privacy protection by infusing synthetic noise on the coordinates of respondents' addresses in a systematic manner. Many geomasking techniques exist, but two examples of common techniques are the donut method and spatial smoothing (Buron & Fontaine, 2018). In the simplest version of the donut method, all geocoded addresses are shifted in a random direction to a distance between a preset minimum and maximum range around the true point. Spatial smoothing involves blending data from local, spatially contiguous areas, thereby creating a weighted average of the values observed in the "neighborhood" of a point (refer to Figure 2 for an example of this technique and the way that smoothing parameters affect the data). Application of spatial smoothing could be as simple as releasing a heat map instead of a choropleth map to blend adjoining areas in the visualization. Geomasking is described in greater detail later in this guide.

Key Concept Highlights

1. Geographical or numerical constraints are set when geomasking to minimize distortion of spatial attributes of microdata or to strengthen privacy protection. Nonresidential areas, such as water, parks, and green areas, can be filtered to ensure that geomasked points do not appear in those areas. A constraint can also limit geomasking displacement to within the same census unit as the original point, which preserves the spatial attributes of respondents as much as possible. Since respondents in areas with low population density are at higher risk of identity disclosure, the extent of displacement can be set inversely proportional to the local population density. Location can be a perfect "key," linking publicly available datasets to otherwise anonymous unassociated datasets with a high degree of confidence.
2. Tools like Google Earth and open data resources make it possible for the public to access information based on location and to "translate" between the different ways that datasets track location.
3. In spatial data, dissimilarity and location can be a powerful combination for a data breach.
4. Nonnested areas can be a critical risk factor for differencing attacks.

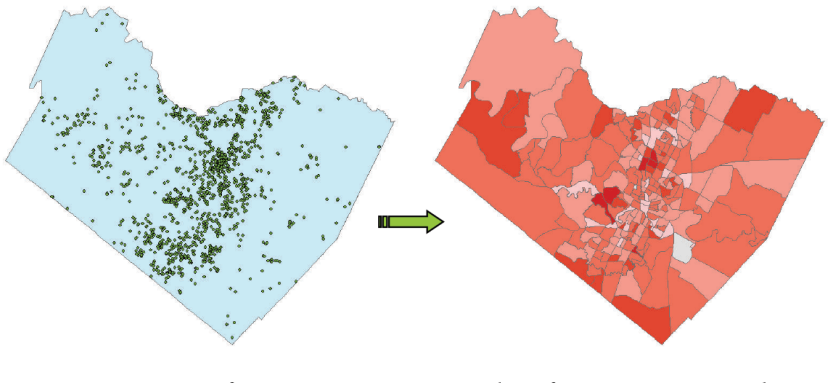
PROCESS OVERVIEW

Release Planning

NSOs should begin planning these three aspects of geographic data release as early in the census process as possible:

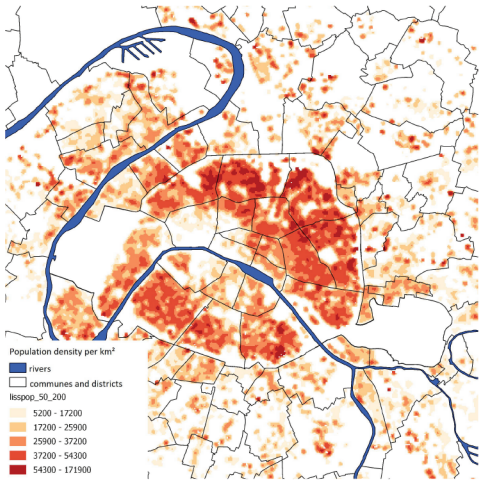
- The specific geographic areas to be released.
- The geographic levels of detail (e.g., province, district, or census block) to be released.
- The form(s) of data to be released (microdata, aggregated polygons, or raster data).

Figure 1.
Spatial Data Aggregation From Individual Points

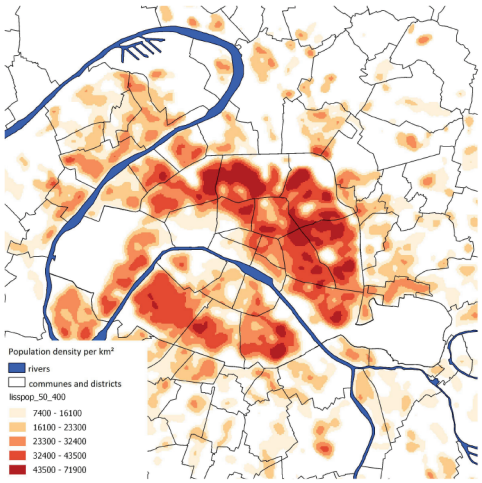


Source: Zandenbergen, 2014.

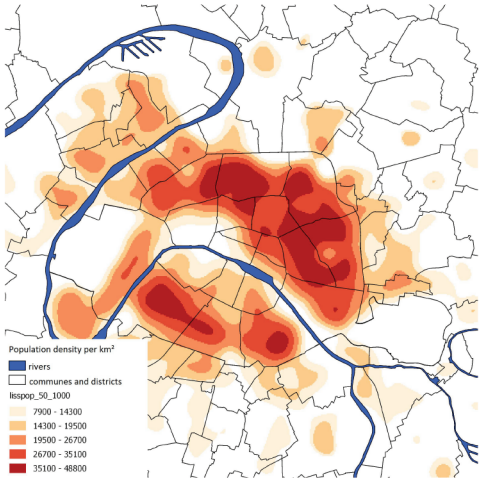
Figure 2.
Three Different Smoothing Radii for Population Density in Paris and Its Suburbs



(a) 200 meter radius



(b) 400 meter radius



(c) 1,000 meter radius

Source: Genebes et al., 2018.

These initial plans are the first steps to take, as they dictate the privacy preservation methods to be used and could involve the NSO releasing fewer products than may be desired.

Evaluate Disclosure Risk

NSOs should begin by reviewing release plans—including past releases—to see if multiple forms exist for the same (or overlapping) data. The release of geomasked point data for an area when a previously released spatially aggregated polygon dataset already exists for the same area decreases the level of privacy protection for both the old and the new product. If so, the overlapping content between the point data and the aggregated data would be flagged as "at risk."

Next, the NSO should check within and between each data product for overlaps between nonnested geographies. Households in overlapping areas would then be flagged as "at risk."

NSOs should then follow the procedures outlined in "Disclosure Avoidance and the Census" (U.S. Census Bureau, 2020), looking for areas where:

- Cells with small counts exist.
- Nonzero counts exist for sensitive groups.
- Different subsets of the results include the same population(s).
- Individuals within a household are already flagged as "at risk."
- Outliers exist within the responses for any variable.

Address the Risk

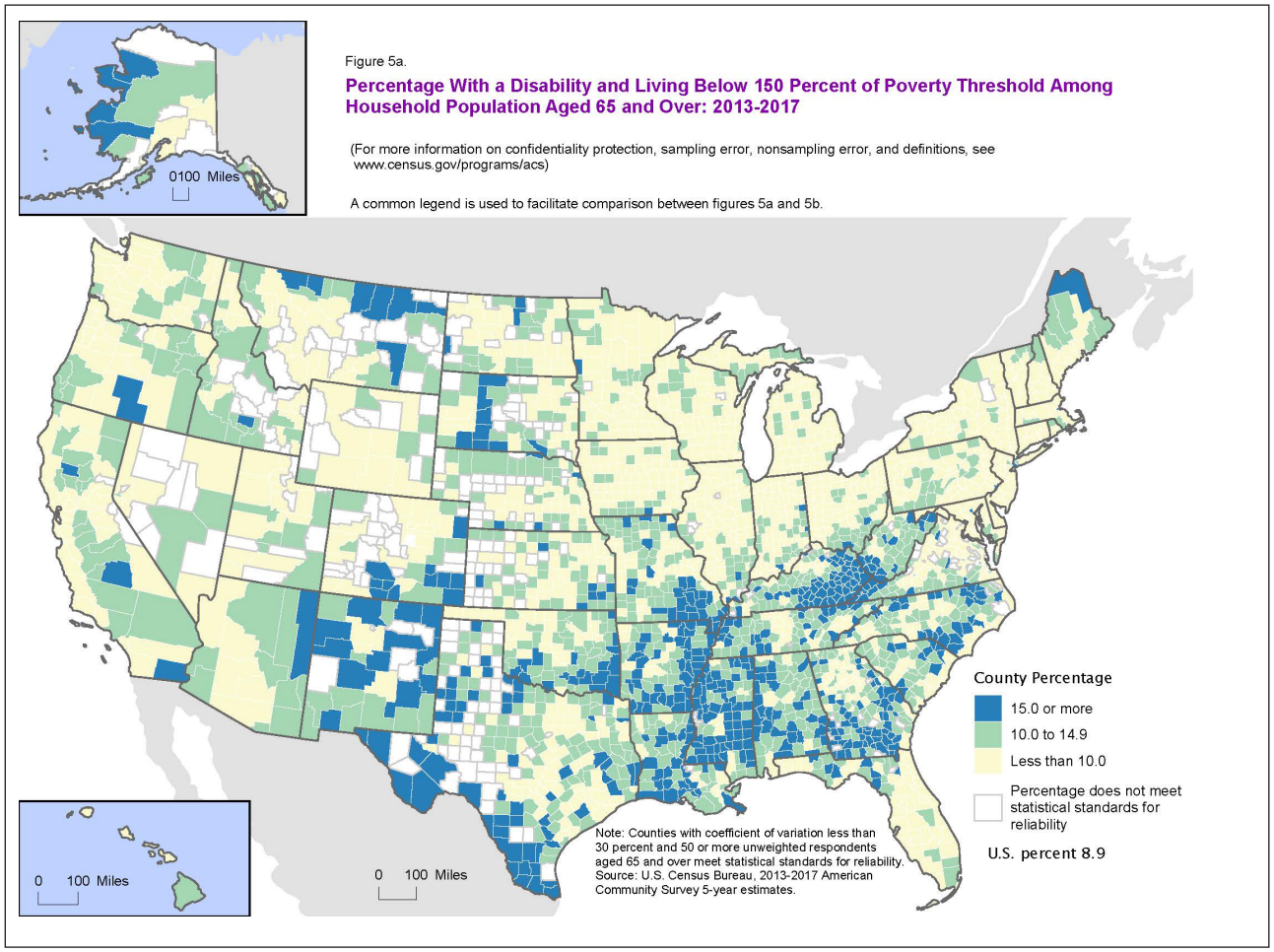
Once the "at risk" datasets, households, and individuals have been flagged, measures may be taken to reduce the risk of disclosure. We cover material specific to geoprivacy concerns in this note, alternative methods that are spatially agnostic—such as controlling access through restricted-access data enclaves—are discussed in "Disclosure Avoidance and the Census" (U.S. Census Bureau, 2020).

It is important to note that no dataset should be released more than once with different privacy protections applied. Two data releases from the same dataset with different methods used can be exploited to de-anonymize respondents.

Suppression

The most straightforward way to protect individual- or household-level data is to simply not release that information, unless the benefit of doing so outweighs any risk of release. Data suppression protects privacy of the respondents by replacing attribute cells in aggregated tables or map objects with a marker synonymous to "no data," when they are below a threshold value to eliminate risk of identity disclosure. This process prevents potential differencing attacks that match areas with low-value cells to other datasets. Refer to Figure 3 for an example where the Census Bureau replaced at-risk polygons with blank polygons indicating "percentage does not meet statistical standards for reliability" when respondent numbers and statistical measures indicated that the population was both at-risk of privacy disclosure, and the data were unreliable for the specific data product. NSOs can suppress entire datasets, omit certain disaggregations (prioritizing

Figure 3.
Example of Census Bureau Product That Excludes At-Risk Flagged Counties



Source: U.S. Census Bureau, <www.census.gov/content/dam/Census/library/visualizations/time-series/demo/older-population/Figure%205%20Population%20Aged%2065%20and%20Over%20with%20a%20Disability%20in%20Poverty.pdf>, 2019.

suppression of those that provide less useful information to the public if released), or omit specific cells.

When suppressing data, NSOs should consider both primary and secondary suppression. Primary suppression involves omitting cells or points where the respondents within were previously flagged. Follow-up secondary suppression involves the suppression of otherwise releasable data where the inclusion of those data would then let the data in the primary suppression cells be de-anonymized. NSOs need to compare data released at different levels of the nested geographic hierarchy to identify cells for secondary suppression (possibly in other datasets) that would otherwise be vulnerable to a differencing attack because of the hierarchical overlap. For further discussion of primary versus secondary suppression, refer to "Disclosure Avoidance and the Census" (U.S. Census Bureau, 2020).

Geomasking

As described above, geomasking measures can be used to statistically infuse synthetic noise on either the coordinates of respondents' addresses or on the attribute data for households or areas. When microdata are released with respondents' coordinates, geomasking can prevent identity disclosure from reverse-geocoding, differencing attack, or a combination of the two. Geomasked coordinates are sufficiently displaced from their original locations and then used to replace the original coordinates in microdata.

The parameters and techniques chosen should be tailored for the location in question. As an example, an algorithm that displaces points randomly within 50–150 meters could be effective in a dense city neighborhood but would likely provide no privacy protection in a rural area.

Spatial Aggregation

Spatial aggregation at coarsened, larger-area census units reduces risk of identity disclosure when an aggregated dataset at a lower-level census unit carries the risk of individual identity disclosure. Aggregated data improves respondents' confidentiality by converting their point-based locations into an area-based census unit. However, if the census units have a small population or a small number of respondents relative to its total population, their identities can be easily inferred from the aggregated data. In those cases, the NSO may aggregate at a coarser spatial unit with a sufficiently large size of population or respondents. The threshold values for aggregation are determined with consideration of sensitivity of data and local population distributions.

Temporal Resolution

When possible, NSOs can adjust the temporal resolution of data being released. If an attacker knows a respondent's location at specific times, it can make it very easy to identify places of work, homes, and places of leisure. Box 3 contains an example of how New York City adjusted temporal resolution in their COVID-19 response.

Box 3.

Case Study: Georeferenced Health Data and the COVID-19 Pandemic

The COVID-19 pandemic sparked awareness of location-based surveillance, quarantining, social distancing, and mobility restriction. Public health authorities conduct contact tracing to slow the spread of COVID-19 in local communities. For precise contact tracing, NSOs and public health authorities collect and combine different georeferenced data resources, including locations of individuals with confirmed positive diagnoses and their movement trajectory records from interviews, social media data, and more. However, the combined georeferenced data have a high risk of identity disclosure that raises concerns about privacy.

When NSOs and public health authorities use nonpublic georeferenced health data, they adopt different methods for data security and privacy protection. For example, South Korea public health authorities follow a privacy protocol that all data are fully anonymized, encrypted, and aggregated at a coarse spatial unit when they conduct contact tracing.

Due to the sensitivity of the health information, the release of the confirmed cases is done with extreme caution on the spatial resolution. New York City releases datasets of confirmed or probable cases, hospitalization cases, and confirmed death cases aggregated at the zip code level. The records are reported daily with a 3-day time lag so identity disclosure can be effectively prevented.

Source: Smith, 2020; NYC Health, 2021; World Bank, 2021.

Check the Results

NSOs should ask three questions in determining the effectiveness of their disclosure control mechanisms.

1. Did those mechanisms sufficiently anonymize the data?
2. Are any distortions in spatial relationships introduced during anonymization minimized?
3. Are the released metadata sufficiently informative without eroding the effectiveness of the anonymization techniques?

Effectiveness of the disclosure controls. To answer the first question—did those mechanisms sufficiently anonymize the data?—spatial k-anonymity may be used to assess anonymity based on the idea that a subject cannot be identified within k-1 other subjects. Spatial k-anonymity is similar to aspatial k-anonymity, however, reverse geocoding is used as the disclosure threat mechanism instead of database record linkage. Even when a geomasked location is displaced from the original location by a substantial distance, it does not solely guarantee confidentiality. If there are a relatively small number of possible residential locations in the area, one of those could be reidentified as the original location of the respondent, and the confidentiality is compromised. This can happen frequently in sparsely populated areas where there are only a few residential locations nearby both geomasked and original locations. It is important to check if the geomasking process provides a sufficiently large number of the residential locations that makes it difficult to reidentify the original location. Before release of the microdata, a minimum threshold value for the number of the closest residential locations can be set to guarantee the level of confidentiality that NSOs want to attain.

Are spatial relationship distortions minimized? To answer the second question—are any distortions in spatial relationships introduced during anonymization minimized?—NSOs must consider and balance the trade-off between spatial distortion and privacy risk. If a higher level of confidentiality is set in the geomasking process (through increasing the displacement distance or increasing the minimum allowable distance to the nearest residential locations nearby the masked locations), the difference between the geomasked and original locations increases, and the quality of the spatial attribute data decreases. This distortion can lead to significant bias in analytical results based on that data. Unfortunately, explicitly assessing for bias requires both the anonymized and nonanonymized datasets, so this is not a test that external researchers can conduct without NSO assistance. However, NSOs could work with researchers to conduct such inquiries for specific research efforts, likely with an approval process and a fee to cover the costs to the NSO.

Are the metadata a risk? To address the third question—are the released metadata sufficiently informative without eroding the effectiveness of the anonymization techniques?—NSOs should release as little metadata as possible on the geomasking and aggregation procedures used in data releases. For example, if an attacker knows that the geomasking method applied used a normal distribution as opposed to a random distribution, the attacker could then more precisely target their attack under the assumption that the true location of a point is likely to lie within a certain range of the jittered point. The release of specific techniques, mask parameters, and the statistical terms that are used substantially decrease the effectiveness of those measures (Kounadi and Leitner, 2014).

CONCLUSION

Spatially aware data are critical for effective public administration in response to both crisis events (Refer to Box 3) and to support general governance. As open data initiatives spread, and improved Geographic Information System (GIS) technologies make spatially aware data more widely available and precise, the risk of geoprivacy data breaches has increased dramatically in recent decades. NSOs are in the advantageous position of having strong statisticians, geographers, and social scientists working closely together. Given appropriate attention to the specific risks and opportunities that geoprivacy measures enable, NSOs are well situated to continue producing data that serve the public good, while balancing the need to maintain respondents' trust through strong geospatial privacy protection measures.

REFERENCES

Antal, L., T. Enderle, and S. Giessing, "Harmonised Protection of Census Data in the ESS: Statistical Disclosure Control Methods for Harmonised Protection of Census Data," Eurostat Centre of Excellence on Statistical Disclosure Control, The Hague, 2017.

Buron, M. L. and M. Fontaine, "14. Confidentiality of Spatial Data," in Loonis, V. and M. de Bellefon, *Handbook of Spatial Analysis: Theory and application with R*, Insee Méthodes No. 131, Eurostat, 2018.

Camenisch, J., S. Fischer-Hubner, and M. Hansen (eds), "Privacy and Identity Management for the Future Internet in the Age of Globalisation (IFIP Advances in Information and Communication Technology)," Springer, 2016.

Genebes, L., Renaud, A., and F. Sémécurbe, "8. Spatial Smoothing," in Loonis, V. and M. de Bellefon, *Handbook of Spatial Analysis: Theory and application with R*, Insee Méthodes No. 131, Eurostat, 2018.

Gutmann, M., K. Witkowski, C. Colyer, J. McFarland O'Rourke, and J. McNalley, "Providing Spatial Data for Secondary Analysis: Issues and current practices related to confidentiality," *Population Research and Policy Review*, 27(6): pp. 639–665, 2009.

Haley, D. F., S. A. Matthews, H. L. F. Cooper, R. Haardörfer, A. A. Adimora, G. M. Wingood, and M. R. Kramer, "Confidentiality Considerations for Use of Social-Spatial Data on the Social Determinants of Health: Sexual and Reproductive Health Case Study," *Social Science and Medicine*, 166: pp. 49–56, 2016.

Kounadi, O. and M. Leitner, "Why Does Geoprivacy Matter? The Scientific Publication of Confidential Data Presented on Maps," *Journal of Empirical Research on Human Research Ethics*, 9(4), pp. 34–45, 2014.

Laaribi, A. and L. Peters, "GIS and the 2020 Census," Esri Press, Redlands, CA, 2019.

Lauger, A., B. Wisniewski, and L. McKenna, "Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research, Research Report Series (Disclosure Avoidance #2014-02)," Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC, 2014.

McKenna, L. and M. Haubach, "Legacy Techniques and Current Research in Disclosure Avoidance at the U.S. Census Bureau," Research and Methodology Directorate, U.S. Census Bureau, Washington, DC, 2019.

NYC Health, COVID-19: Latest Data - NYC Health, <https://www1.nyc.gov/site/doh/covid/covid-19-data.page#epicurve>, accessed on March 11, 2021.

Pascale, J., D. K. Willimack, N. Bates, J. F. Lineback, and P. C. Beatty, "Issue Paper on Disclosure Review for Information Products with Qualitative Research Findings," Research and Methodology Directorate, U.S. Census Bureau, Washington, DC, 2020.

Smith, C. D. and J. Mennis, "Incorporating Geographic Information Science and Technology in Response to the COVID-19 Pandemic," Preventing Chronic Disease, 17, E58, 2020.

United Nations Statistics Division, "Principles and Recommendations for Population and Housing Censuses, Revision 3," United Nations Publications, New York, NY, 2015.

United States Census Bureau, "Disclosure Avoidance and the Census," Select Topics in International Censuses, <www.census.gov/content/dam/Census/library/working-papers/2020/demo/disclosure_avoidance_and_the_census_brief.pdf>, 2020.

United States Census Bureau, "Figure 5: Population Aged 65 and Over With a Disability in Poverty, Select Maps on the Population 65 and Older in the United States by County: 2013–2017," <www.census.gov/content/dam/Census/library/visualizations/time-series/demo/older-population/Figure%205%20Population%20Aged%2065%20and%20Over%20with%20a%20Disability%20in%20Poverty.pdf>, 2019.

World Bank, "The Role of Geospatial Information in Confronting COVID-19 – Learning From Korea," <https://blogs.worldbank.org/eastasiapacific/role-geospatial-information-confronting-covid-19-learning-korea>, accessed on March 11, 2021.

Zandenbergen, P. A., "Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data," *Advances in Medicine*, 2014.

Zayatz, L., "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update," Research Report Series (Statistics #2005-06), Statistical Research Division, U.S. Census Bureau, Washington, DC, 2005.

The Select Topics in International Censuses (STIC) series is published by International Programs in the U.S. Census Bureau's Population Division. The United States Agency for International Development sponsors production of the STIC series, as well as the bilateral support to statistical organizations that inform authors' expertise. The United Nations Population Fund collaborates on content and dissemination, ensuring that the STIC series reaches a wider audience.